Retrieving Customary Web Language to Assist Writers

Benno Stein, Martin Potthast, and Martin Trenkmann

Given a phrase of text, like this one.

Is there a more customary alternative?
What comes next, before, or in between?
Is the word order correct?
Is the preposition properly used?
Should there be a comma?
In short: *How would others write it?*

What if you‘re not a native speaker?

Even if so...

The Web is a large resource of modern English.
It can be exploited to retrieve *customary* language.

To date the Google *n*-gram corpus is the largest text sample of its kind; it covers the Web of 2006.
We index this collection to realize the NETSPEAK writing assistant.

Example queries:

```
a ? of text
given * like this
{ like this one }
phrase ? text
? like this
```

The n-gram corpus before and after post-processing:

Corpus Subset	Original Corpus		Case Reduction	Vocabulary Filtering
	n-grams	Size		
1-gram	13 588 391	177.0 MB	81.34 %	3.75 %
2-gram	314 843 401	5.0 GB	75.12 %	43.26 %
3-gram	977 069 902	19.0 GB	83.24 %	48.65 %
4-gram	1 313 818 354	30.5 GB	90.27 %	49.54 %
5-gram	1 176 470 663	32.1 GB	94.13 %	47.16 %
Σ	3 354 253 200	77.9 GB	88.37 %	54.20 %

EBNF grammar of the query language:

Production Rule		
query	=	$\{\text{word} \mid \text{wildcard}\}_1^5$
word	=	(["'"] (letter {alpha})) | ","
letter	=	"a" |...| "z" | "A" |...| "Z"
alpha	=	letter | "0" |...| "9"
wildcard	=	"?" | "*" | synonyms | multiset
synonyms	=	"~" word
multiset	=	"{" word {word} "}"

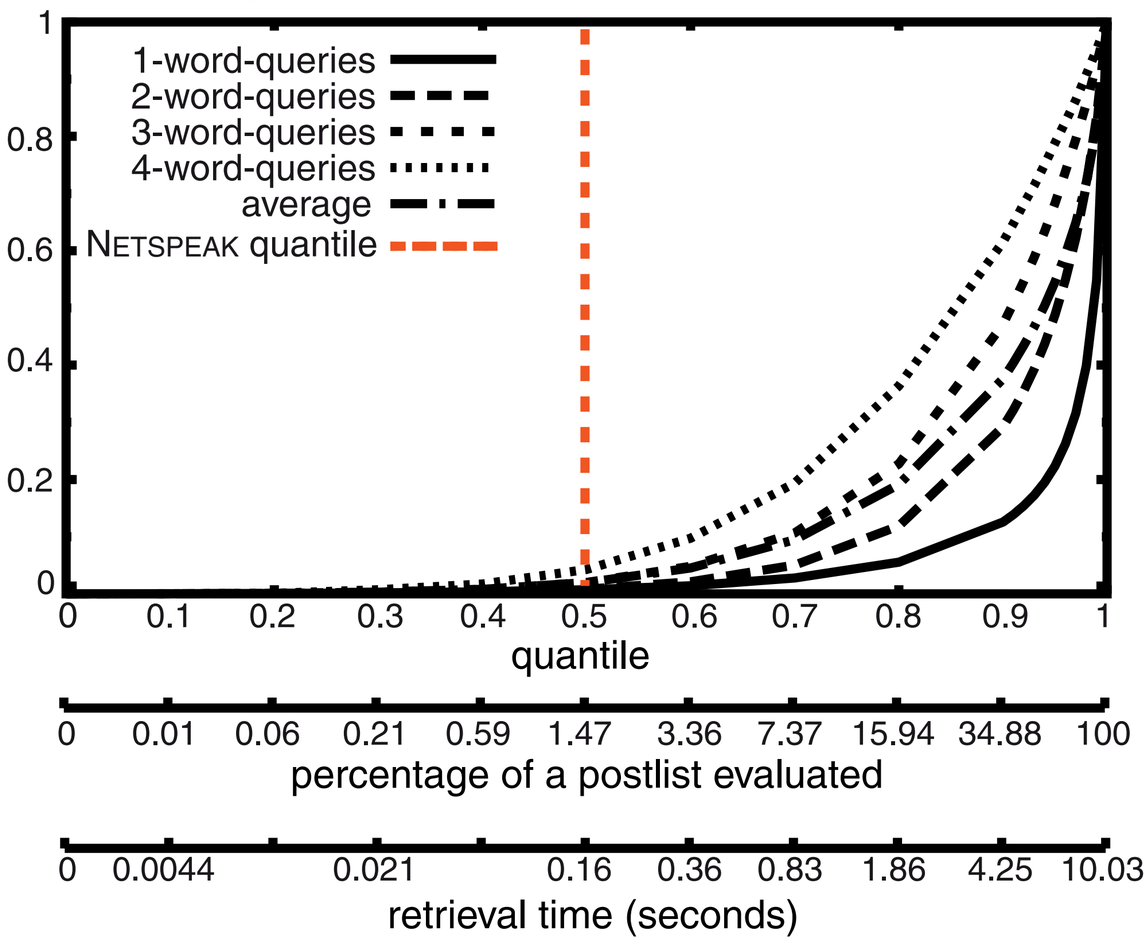
Our index implements a top-*k* probabilistic retrieval strategy which allows to prune the search.

Key concepts are sorted postlists and quantile entries that divide the postlist frequency distribution.
Results are thus retrieved by decreasing relevance.

We evaluate the recall of frequent *n*-grams, when pruning the search at a given quantile.

The retrieval performance is measured with regard to macro- / micro-averaged recall.
The plots indicate a speedup of factor 68, compared to an unpruned search.

Macro-averaged recall



Micro-averaged recall